A note on Regression Estimates in Stratified Sampling

Kim Ik-chan, Yang Sung-ho *

層化標本抽出에서의 回歸推定値에 관한 小考

金益賛, 梁成豪*

1. Introduction

The linear regression estimate can be designed to increase precision by the use of an auxiliary variate x_i that is correlated with y_i like the ratio estimate. When the relation between y_i and x_i is examined, it may be found that although the relation is approximately linear, the line does not go through the origin.

This suggests an estimate based on the linear regression of y_i on x_i rather than on the ratio of the two variables. We suppose that y_i and x_i are each obtained for every unit in the sample and that the population mean $\bar{X}$ of the x_i is known.

The linear regression estimate of $\bar{Y}$, the population mean of the y_i, is

$$\bar{Y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) \qquad (1.1)$$

Where the subscript lr denotes linear regression and b is an estimate of the change in y when x is increased by unity. The rationale of this estimate is that if $\bar{x}$ is below average we should expect $\bar{y}$ also to be below average by an amount $b(\bar{X} - \bar{x})$ because of the regression of y_i on x_i. For an estimate of the population total Y, we take $\hat{Y}_{lr} = N\bar{y}_{lr}$.

2. Notation

In stratified sampling the population of N units is first divided into subpopulations of $N_1, N_2, \cdots\cdots, N_L$ units, respectively.

These subpopulations are nonoverlapping, and together they comprise the whole of the population, so that $N_1 + N_2 + \cdots\cdots + N_L = N$.

The subpopulations are called strata. The sample size within the strata are denoted by $n_1, n_2 \cdots\cdots n_L$, respectively. The suffix h denote the stratum and i the unit within the stratum. The following symbols all refer to stratum h.

N_h : total number of units

師範大學 助教授, 師範大學 助教授*

n_h : number of units in sample
y_{h_i} : value obtained for ith unit
$W_h = N_h/N$: stratum weight
$f_h = n_h/N_h$: sampling fraction in the stratum
$f = h/N$: sampling fraction
$\bar{Y}_h = \sum_{i=1}^{N_h} y_{h_i}/N_h$: true mean
$\bar{y}_h = \sum_{i=1}^{n_h} y_{h_i}/N_h$: sample mean
$S^2_h = \sum_{i=1}^{N_h} (y_{h_i} - \bar{y}_h)^2/(N_h - 1)$: true variance

3. Theorems

There are two ways in which a regression estimate can be made in stratified random sampling. One is to make a separate regression estimate $\bar{y}_{lrs}$, computed for each stratum mean, that is.

$$\bar{y}_{lrh} = \bar{y}_h + b_h(\bar{X}_h - \bar{x}_h) \tag{3.1}$$

then, with $W_h = N_h/N$,

$$\bar{y}_{lrs} = \sum_h W_h \bar{y}_{lrh} \tag{3.2}$$

An alternative combined regression estimate, $\bar{y}_{lrc}$ is derived by combining estimates in stratified sampling. To compute $\bar{y}_{lrc}$, we first find

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \qquad \bar{x}_{st} = \sum_h W_h \bar{x}_h. \tag{3.3}$$

Then

$$\bar{y}_{lrs} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st}) \tag{3.4}$$

Preliminary 1.

In simple random sampling, in which b_0 is a preassigned constant, the linear regression estimate

$$\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x}) \tag{3.5}$$

is unbiased, with variance

$$V(\bar{y}_{lr}) = \frac{1-f}{n}(S_y^2 - 2b_0 S_{yx} + b_0^2 S_x^2) \tag{3.6}$$

Proof

See [Cochran]

Preliminary 2.

The value of b_0 that minizes $V(\bar{y}_{lr})$ is

$$b_0 = B = S_{yx}/S_x^2 = \sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X}) / \sum_{i=1}^{N}(x_i - \bar{X})^2 \tag{3.7}$$

And the minimum variance is

$$V_{min}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2(1-\rho^2) \tag{3.8}$$

where ρ is the population correlation coefficient between y and x.

Proof

see [Cochran]

Theorem 1.

The linear reqression estimate $\bar{y}_{lrs}$ (s for seperate), (3.2) is unbiased estimate of $\bar{Y}$, with variance

$$V(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2 - 2b_h S_{yxh} + b_h^2 S_{xh}^2) \tag{3.9}$$

Proof

Each stratum mean $\bar{y}_{lrh}$ is the sample mean of the quantities $y_{h_i} - b_h(x_{h_i} - \bar{X})$. Hence by Preliminary 1

$$E(\bar{y}_{lrs}) = E\sum_h W_h \bar{y}_{lrh} = \sum W_h \bar{Y}_h = \frac{\sum N_h \bar{Y}_h}{N} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{N_h} y_{hi}}{N} = \bar{Y} \tag{3.10}$$

And

$$V(\bar{y}_{lrs}) = V(\sum W_h \bar{y}_{lrh}) = \sum W_h^2 V(\bar{y}_{lrh}) \tag{3.11}$$

On the other hand

$$V(\bar{y}_{lrh}) = \frac{1-f_h}{n_h} \cdot \frac{\sum[(y_{hi} - \bar{Y}_h) - b_h(x_{h_i} - \bar{X}_h)]^2}{N-1}$$

$$= \frac{1-f_h}{n_h}(S_{yh}^2 - 2b_h S_{yxh} + b_h^2 S_{xh}^2) \tag{3.12}$$

Sustituting (3.12) into (3.11)

$$V(\bar{y}_{lrs}) = \sum \frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2 - 2b_h S_{yxh} + b_h^2 S_{xh}^2) \tag{3.13}$$

Theorem 2.

$V(\bar{y}_{lrs})$ is minimized when $b_h = B_h$, the true regression coefficient in stratum h.

And the minimum value of the variance is

$$V_{min}(\bar{y}_{lrs})=\sum\frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2-\frac{S_{yxh}^2}{S_{xh}^2}) \quad (3.14)$$

Proof

By the Preliminary 2. $V(\bar{y}_{lrs})$ is minimized

when $b_h=B_h=\frac{S_{yxh}}{S_{xh}^2}$ (3.15)

By partially differentiation (3.13) with respect to b_h and substituting (3.15) into $V(\bar{y}_{lrs})$ then

$$V_{min}(\bar{y}_{lrs})=\sum_h\frac{W_h{}^2(1-f_h)}{n_h}(S_{yh}^2-\frac{S_{yxh}^2}{S_{xh}^2})$$

Theorem 3

The combined regression estimate $\bar{y}_{lrc}$ is an unbiased estimate of $\bar{Y}$ with variance

$$V(\bar{y}_{lrc})=\sum_h\frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2-2bS_{yxl}+b^2S_{xh}^2) \quad (3.16)$$

Proof

By Preliminary 1

$$E(\bar{y}_{lrc})=E[\bar{y}_{st}+b(\bar{X}-x_{st})]$$
$$=E(\sum_h W_h\bar{y}_h)+E[b(\bar{X}-\sum W_h\bar{x}_h)]$$
$$=\bar{Y} \quad (3.17)$$

Since $\bar{y}_{lrc}$ is the usual estimate from the stratified sample for the variate $y_{h_i}+b(\bar{X}-x_{h_i})$. and the variance of the estimate $\bar{y}_{st}$ is

$$V(\bar{y}_{st})=\frac{1}{N^2}\sum_{h=1}^{L}N_h(N_h-n_h)\frac{W_h^2}{n_h}$$
$$=\sum_{h=1}^{L}W_h^2\frac{S_h^2}{n_h}(1-f_h) \quad (3.18)$$

hence

$$V(\bar{y}_{lrc})=\sum\frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2-2bS_{yxh}+b^2S_{xh}^2)$$

Theorem 4.

The value of b that minimizes the variance of (3.16) is

$$B_c=\sum_h\frac{W_h^2(1-f_h)S_{yxh}}{n_h}/\sum_h\frac{W_h^2(1-f_h)S_{xh}^2}{n_h} \quad (3.19)$$

Proof

From (3.16)

$$\frac{\partial V(\bar{y}_{lrc})}{\partial b}=\sum\frac{W_h^2(1-f_h)}{n_h}(-2S_{yxh}+2S_{xh}b)$$
$$=0$$

then $b=\sum\frac{W_h^2(1-f_h)S_{yzx_h}}{n_h}/\sum\frac{W_h^2(1-f_h)S_{xh}^2}{n_h}$

is the minimized variance.

hence $B_c=\sum\frac{W_h^2(1-f_h)S_{yxh}}{n_h}/\sum\frac{W_h^2(1-f_h)S_{xh}^2}{n_h}$

Theorem 5.

$$V_{min}(\bar{y}_{lrc})-V_{min}(\bar{y}_{lrs})=\sum a_h(B_h-B_c)^2 \quad (3.20)$$

where $a_h=\frac{W_h^2(1-f_h)}{n_h}S_{xh}^2$

Proof

$$V_{min}(\bar{y}_{lrs})=\sum_h\frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2-2B_cS_{yxh}+B_c^2S_{xh}^2)$$

$$V_{min}=(\bar{y}_{lrs})=\sum_h\frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2-\frac{S_{yxh}^2}{S_{xh}^2})$$

$$V_{min}(\bar{y}_{lrc})-V_{min}(\bar{y}_{lrs})=\sum_h\frac{W_h^2(1-fh)}{n_h}(-2B_cS_{yxh}+B_c^2S_{xh}^2+\frac{S_{yxh}^2}{S_{xh}^2})$$

$$=\sum a_hB_h^2+\sum a_hB_c^2-2\sum a_hB_cB_h$$
$$=\sum a_h(B_h^2+B_c^2-2B_cB_h)$$
$$=\sum a_h(B_h-B_c)^2$$

This result shows that with the optimum choices the separate estimate has a smaller variance than the combined estimate unless B_h is the same in all strata.

In comparing of the two types of estimate, if we are confident that the regressions are linear and B_h appears to be roughly the same in all strata, the combined estimate is to be preferred. If the regressions appear linear, so that the danger of bias seems small, but B_h seems to vary markly from stratum to stratum, the separate estimate is advisable. If there is some curvilinearity in the regressions when a linear regression estimate is used, the combined estimate is safer unless the samples are large in all strata.

Literature cited

Cochran, W. G., 1977. Sampling Techniques. John Wiley & sons, Inc. New York. 3rd ed. p. 89–204.

Draper, N. R., 1981. Appried Regression Analysis John Wiley & sons, Inc. New York. 2nd ed. p. 117–124.

Park sung-hyun. 1984. 回歸分析. 大英社. p. 137–170.

Taro Yamane. 1967. Elementary Sampling Theory. Prentice–hall. Inc. Englewood Cliffs. N. J. p. 121–158.

國 文 抄 錄

線型回歸推定은 精度를 높인다. 특히 層化 標本抽出에서의 回歸推定에는 두가지 方法이 있다. 즉 분리된 回歸推定과 결합된 回歸推定 方法이다. 결합형 推定値는 모든 層別에서 그 係數가 동일한 경우에 분리형에서는 層別간 현저한 변화가 있는 경우에 유용하게 적용된다. 이들 두 형태의 回歸推定値에 관한 推定量과 最少分散値 및 偏倚差에 관한 정리를 고찰하였다.